Historic, Archive Document

Do not assume content reflects current scientific knowledge, policies, or practices.

United States
Department of
Agriculture

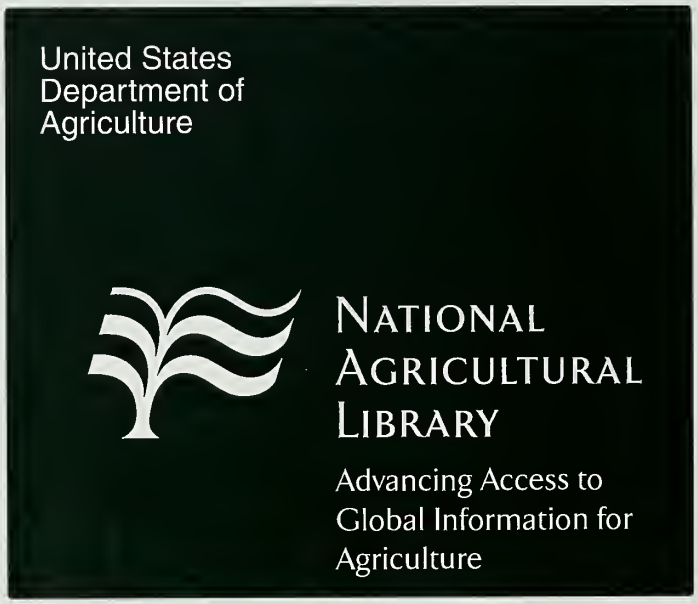
Agricultural
Research
Service

Program Aid 1500

March 1994

The USDA-ARS Plant Genome Research Program

1
Ag84Pro
no. 1500



Agricultural Research Service

The United States Department of Agriculture (USDA) prohibits discrimination in its programs on the basis of race, color, national origin, sex, religion, age, disability, political beliefs and marital or familial status. (Not all prohibited bases apply to all programs). Persons with disabilities who require alternative means for communication of program information (braille, large print, audiotape, etc.) should contact the USDA Office of Communications at (202) 720-5881 (voice) or (202) 720-7808 (TDD).

To file a complaint, write the Secretary of Agriculture, U.S. Department of Agriculture, Washington, D.C., 20250, or call (202) 720-7327 (voice) or (202) 720-1127 (TDD). USDA is an equal opportunity employer.

Foreword

A new generation of scientific tools is now available for the study of plant genes. The capability now exists to rapidly locate genes on chromosomes, to isolate genes from plants, to study their functioning at the molecular level, to modify genes, and to reintroduce them into living organisms. These new tools have revolutionized the science of genetics, and they are already yielding a great new flood of information about the way in which individual genes control specific plant characters.

Many of us will feel the impact of these discoveries, but none of us more directly than the policymakers who will come to grips with key issues affecting the course of future scientific endeavor. How can the Federal research sector most wisely and expeditiously organize the new biotechnological discoveries that are continuously occurring and implement them, integrating new possibilities into existing programs that work to solve profound and longstanding problems in agriculture? To do so, one needs a specialized vantage on the state of today's research, which stands at a remarkable threshold.

For while terms such as "gene shuttling "and "transfer" have entered the popular vocabulary, the scientific framework that underpins the jargon cannot be encapsulated quite as concisely. After all, plants contain thousands of genes, and each plant's assemblage of genes is organized on chromosomes in a complex genome. The new tools of molecular genetics allow us to search for genes of agricultural importance, study the structure and organization of these genomes, and construct detailed maps of the various desired components of the genome. The way in which genomes can be mapped and the benefits to be derived from such studies are the subjects at hand.

This document serves to organize the subject of plant genome mapping around two objectives: to enhance the layperson's comprehension, and to add context and specifics to the technical reader's grasp. It provides a nontechnical ledger on the left side of the page that invites rapid scanning, while the body of text appears in a right-hand column. To help readers assimilate the inevitable new terminology, boxed definitions of certain terms are included at the bottom of the page; a comprehensive glossary appears afterward.

Jerome P. Miksche, Director
USDA-ARS Plant Genome Research Program

Support for preparation of this publication was provided to Dr. Gary Kochert, Botany Department, University of Georgia, by the USDA-ARS Plant Genome Research Program.

Additional copies may be obtained by writing to Jerome P. Miksche, USDA-ARS, Room 331C, Building 005, BARC-West, Beltsville, MD 20705.

Contents

The USDA-ARS Plant Genome Research Program

Basic and applied biotechnology: how it began, where agricultural research is taking it, and how scientists plan to get there.

It's in the Genes

To begin with, there were plant traits. Traits that were of the highest importance to farmers, like hardiness, high yield, or sweet-tasting fruit.

Agriculture and genetics were born when people began to save some of the seeds they collected for food and to plant them in later seasons or in more desired habitats. Two crucial observations must have been made early in this process. The first was that variation was present in plants used by humans. Variation in characters such as seed size or number, flavor of edible parts, and earliness to harvest would have been of great interest to early farmers. The second crucial observation was that characteristics tend to be inherited, and crops could be improved by selecting seeds from favorable plant types for later propagation.

How could a farmer know if these traits would pass from one generation to the next? Would next year's seed bear true? Or was plant heredity merely a roll of the dice?

The first genetic experiments were largely empirical and observational. Selection for superior plant types could be done, and was very effectively done, without any understanding of the basic mechanisms by which characters are transmitted from one generation to the next or the way in which characters are stored and expressed.

Scientific progress is often limited by the tools available to practitioners. Before fundamental progress could be made toward another level of genetic understanding, new tools had to be developed. Although the intellectual capabilities of scientists have not increased, new techniques or instrumentation periodically provide the opportunity for sudden advancement in scientific understanding.

The ability to observe samples at increased magnification in microscopes or telescopes was one such advance. This extension of our visual acuity brought about fundamental changes in how we look at the world and in our theories of how it works. Microscopic observation of cells, cell division, and chromosome behavior was essential to the great advances made in genetics in the first half of the 20th century. Our present ideas about genetic maps and gene transmission all trace their origin to this seminal period.

Checking for desirable traits in wheat plants carrying wheat-rye chromosomal translocations.

Character—*An observable feature (phenotype) of the fully developed organism; for example, red flowers, dwarf plants.*

Variation—*Heritable and nonheritable differences in structures at the cell, individual, and among individual levels.*

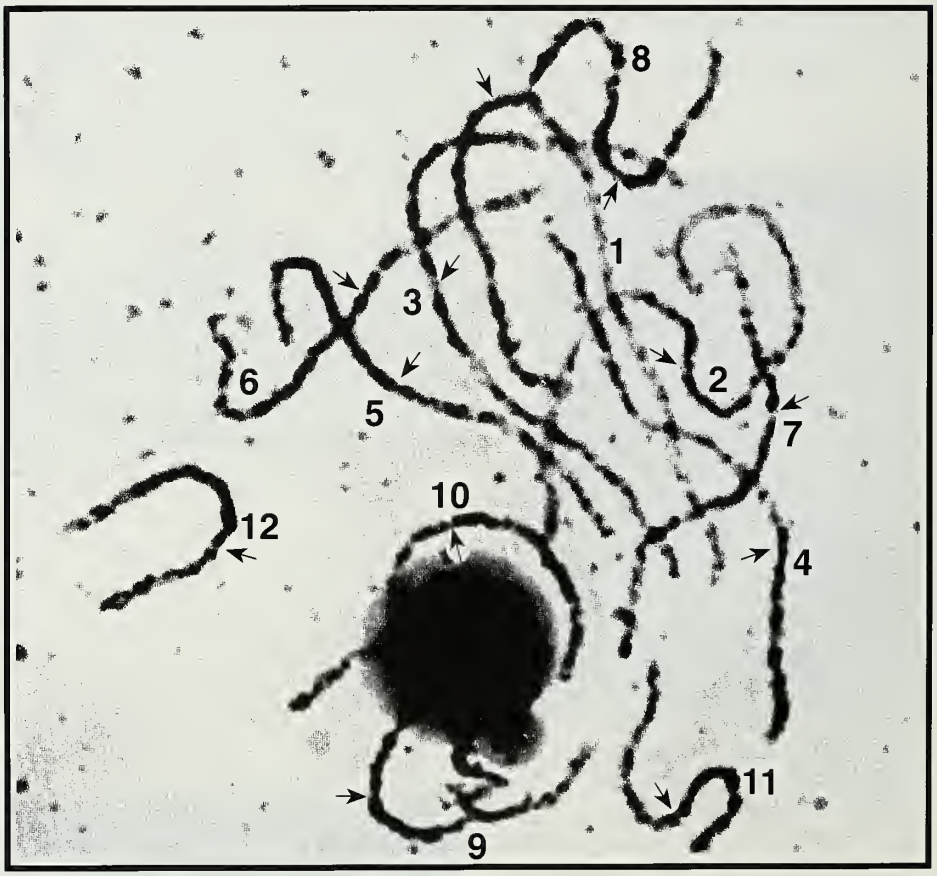
Reviewing a projection of tobacco pollen from transgenic plants.

It was a great step forward for the study of genetics when Gregor Mendel, using numerically compiled observations, discovered some basic patterns to trait inheritance.

But, scientists wondered, what caused these patterns to be? There must be physical mechanisms responsible for heredity. As the technology of instruments advanced, evidence pointed to structures within the cell itself.

One of the most significant conceptual advances in science was made in the 19th century by Gregor Mendel, who postulated that plant characters are controlled by discrete fundamental units, which we now call genes. Studying patterns of inheritance in pea blossoms, Mendel showed that genes retain their identity during transmission from one generation to another and do not blend. Geneticists took up the search for the location of these genes in cells, their chemical composition, and their mode of action in the first decades of the 20th century after the rediscovery of Mendel's results.

Researchers, using the newly improved microscopes available in the late 19th and early 20th centuries, noted that cells contained brightly staining, thread-like objects in their nucleus. Remarkably, each of these chromosomes, as they were named, had the ability to replicate itself, and the subsequent copies were precisely distributed during cell division; each daughter cell received an exact copy of the parent's chromosomes.



A microscopic view of the chromosomes of rice. Twelve chromosome pairs can be seen in this pachytene view. The large spherical object is the nucleolus.

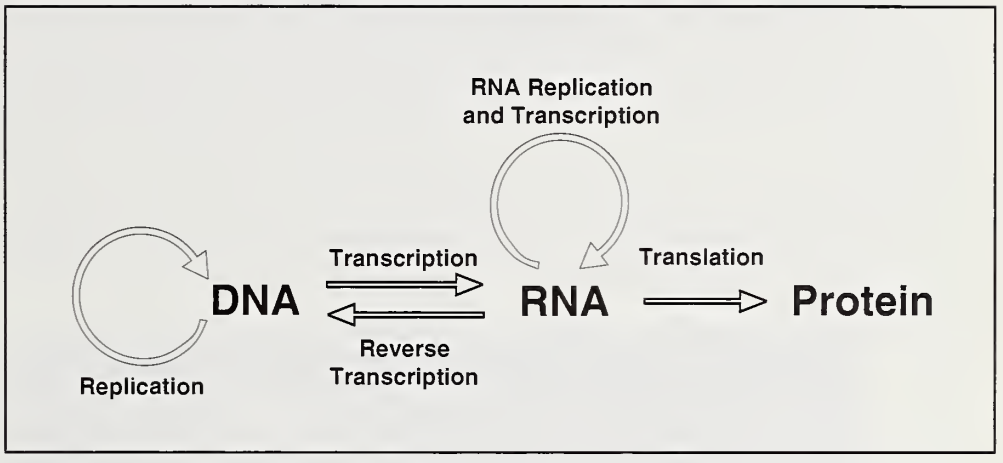
Such precise distribution of a cell component must be of fundamental importance to the cell, it was reasoned, and chromosomes were soon postulated to be the carriers of the genes. After scientists conducted detailed

studies of the behavior of chromosomes during cell division—and especially during the specialized set of cell divisions that give rise to sperm and eggs—the idea that genes must be associated with chromosomes was confirmed.

New combinations of genes could be shown to arise by breakage and reunion of chromosomes, and genes were thus demonstrated to have definite locations on chromosomes. It followed that genes must be linearly arranged in single file on chromosomes and that it would be possible to make a map of the location of genes relative to one another.

The Composition of Genes

Advances in cytochemistry, cell fractionation, and biochemical analysis that began in the 1940's made it possible to analyze the chemical nature of chromosomes. Proteins and DNA were found to be universally present, and there was considerable controversy about which component might be the carrier of genetic information. Experiments demonstrated that genetic



The "central dogma" of molecular genetics: the genetic code in DNA finds its expression in proteins.

Proteins—*organic molecules that are rich in nitrogen and are composed of polypeptides made from the combination of 20 amino acids.*

Cytochemistry—*identification and localization of chemical components at the cell and subcellular levels.*

Cell fractionation—*a method to break down cells into their separate constituents and subcellular structures for biochemical analysis.*

Interpreting a gel map of wheat DNA samples.

characters could be transferred from one bacterium to another by exposing them to the appropriate DNA. This conclusively showed that DNA must be the material of which genes are constructed.

So cells contained chromosomes, and it was postulated that chromosomes contained genes. But how did these structures operate? What was their biochemical function? Nobelists Watson and Crick made an enormous breakthrough when, in the early 1950's, they succeeded in describing the structure of DNA.

DNA is a deceptively simple molecule. It contains four basic nucleotide components arranged as nucleotide pairs in linear order on two parallel strands wound around in the famous double helix. The most significant variability in its structure is simply the order in which the four components occur in the various parts of the molecule. Everything an organism is or potentially can do is inherent in the order of those four components.

Even though genes are made of DNA, it's not DNA that carries out the function of genes. Instead the DNA of a gene contains a code, in the form of its nucleotide sequence, that specifies the construction of a protein that will mediate the action of the gene. Proteins are formed by expression of this genetic code contained in genes, and gene expression is precisely controlled so that genes can be expressed in the proper temporal and spatial context.

No one knows the precise number of genes present in any complex organism, but crop plants are estimated to contain about 50,000 genes. A gene is a region of the DNA molecule where the proper sequence of nucleotides exists to code for the gene product, which is ultimately almost always a protein. The amount of DNA needed to code for a typical gene product appears to be fewer than 1,000 nucleotide pairs. For 50,000 genes, about 10^7 nucleotides would be sufficient, and plants with the smallest known genomes, such as *Arabidopsis*, have genomes of about this size. Most crop plants, however, have genomes of 10^9 to 10^{10} nucleotide pairs. So there's a great deal of extra (some of it repetitive) DNA in plant genomes. This extra DNA complicates the problem of characterizing genomes and locating genes.

If then, DNA is a configuration of paired nucleotides, scientists reasoned, it would be valuable to spell out relationships between these configurations associated with different traits, and for different species. Thus was born the concept of gene mapping.

Mapping the location of genes on chromosomes has been an important research activity for many years. Chromosome maps are of two basic types: (1) genetic maps, based on recombination values, or the frequency with which chromosomes break or unite to form new combinations of genes during gamete formation, and (2) physical maps, which are based on the nucleotide sequence of the DNA. Until recently, it was only possible to construct genetic maps. Construction of such genetic maps has been a very time-consuming and expensive operation, and genetic maps are available for only a few organisms.

Each existing map represents many years of labor by scientists in several different laboratories. For example, genetic mapping in rice started in the 1920's. In the 70 ensuing years, only about 100 genes have been located on the chromosomes. Molecular genetic techniques have made it possible to construct genetic maps with a fraction of the former expense and labor. In addition, the techniques necessary to develop physical maps are rapidly being developed.

Genetic Maps

Mendel's laws served as a starting point for early efforts to describe genes and their behavior.

Genetic maps are constructed by observing the segregation patterns of genes in the progeny or offspring derived from crossing two parental organisms with contrasting genes. For example, one might cross a plant with yellow flowers and rough seeds to another plant with white flowers and smooth seeds. If the flower color character and the seed coat character are each controlled by single genes, two sorts of progeny might occur: parental types, with the same flower and seed character as the parents, or recombinant types, which would have yellow flowers and smooth seeds or white flowers and rough seeds.

If the progeny are mostly parental, the genes are said to be "linked," meaning they are close together on a chromosome. If the progeny are a more or less equal mixture of parental type and nonparentals, the genes are said to be unlinked, meaning they are far apart on a chromosome or perhaps on separate chromosomes.

The assumption underlying genetic mapping is that the farther apart genes are from one another on a chromosome, the more often they will be able to break apart and form new (nonparental) combinations with other genes in the process called recombination. Genetic map distances, which are based on recombination, cannot be directly related to distances in nucleotide pairs on DNA.

A genetic map shows the relative position of genes on a chromosome and the relative distance between genes. Genetic maps are difficult to construct for several reasons. Before any mapping cross can be made, parents with contrasting genes must be located. To map genes controlling flower color, for

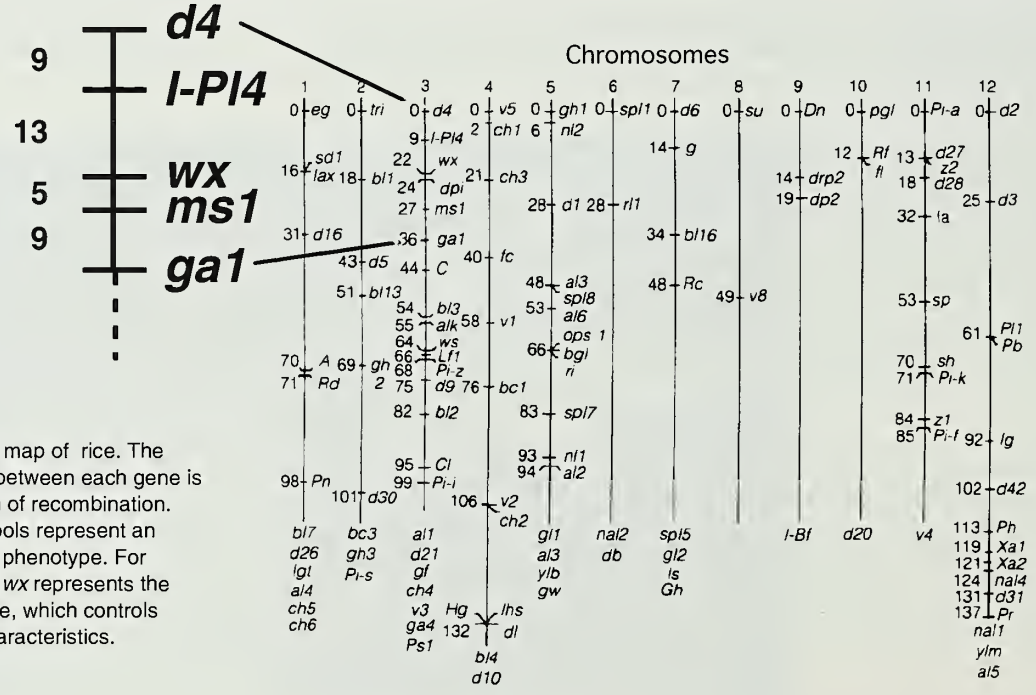
Genetically-engineered potato tubers are readied for field trials.

example, parent plants must be located that have differently colored flowers. Sometimes these can't be found in nature and must be artificially created by mutagenesis.

Only a small number of contrasting or variable gene sets can be formed in any one parent pair, so to increase that number, many different crosses must be made. Also, telling apart parental and nonparental progeny from a cross for a certain gene depends on observing the external morphological or physiological effect of the gene on the organism. There is no way to "score" the gene directly at the DNA level. For flower color genes, for example, it would be necessary to grow the progeny plants to a size suitable for flowering and then to observe and score, or denote, the color of the flowers (phenotype) of the progeny. It is obvious that this requirement would greatly inhibit the study of plants, such as trees, which require years of growth before flowering. So genetic maps of conventional genes are available for only a few plants, and many of these are low-resolution maps with only a few marker genes.

RFLP Maps

In recent years, developments in DNA cloning have enabled scientists to more quickly and efficiently construct genetic maps. Previously, the only way to score or detect the segregation of chromosomal DNA was to infer it indirectly by observing the end result of the action of genes contained on the



A genetic map of rice. The distance between each gene is a function of recombination. The symbols represent an observed phenotype. For example, *wx* represents the waxy gene, which controls starch characteristics.

DNA. It is now possible to ascertain the segregation of pieces of chromosomal DNA directly and to construct genetic maps based on these DNA differences.

Differences in DNA between potential parents in a genetic cross can now be detected in several ways. One of these uses a class of enzymes called restriction endonucleases. They have the ability to search out a distinct sequence of nucleotides in the DNA and cut the DNA molecule at that location. Since the size of each individual fragment produced depends on the nucleotide sequence of the DNA, different sizes (or lengths) of restriction fragments are typically produced when different organisms are tested. Such differences are called restriction fragment length polymorphisms (RFLP's).

When genomic DNA of a crop plant is digested with a restriction enzyme, an enormous number of DNA fragments (restriction fragments) are produced. To detect RFLP's, it's necessary to identify individual restriction fragments from the complex mixtures produced by digestion of a complex genome. This was first accomplished by the use of cloned DNA fragments as specific probes. Nucleic acid hybridization was used to identify individual restriction fragments that could hybridize to the probe.

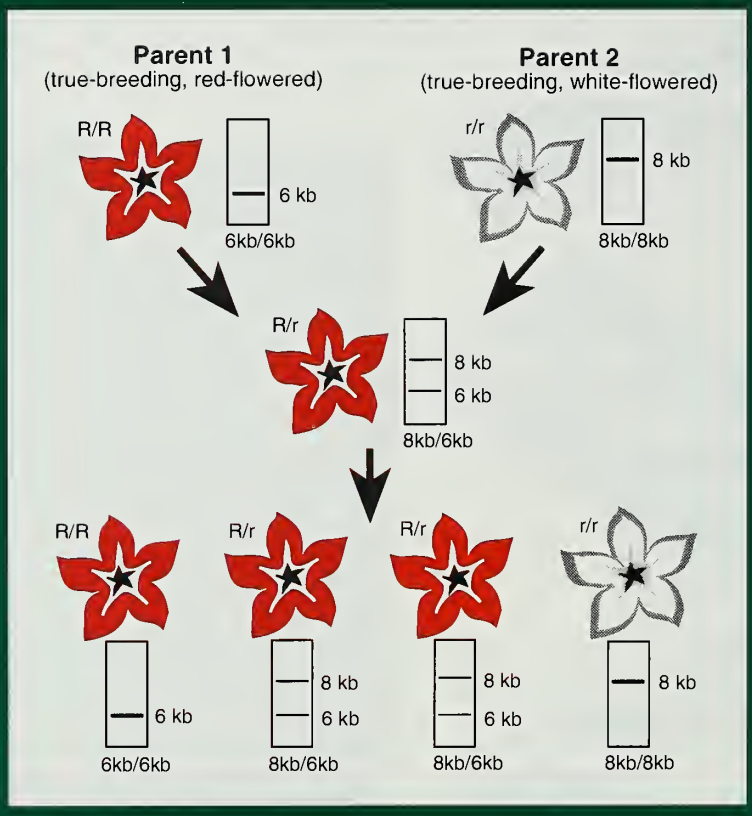
In practice, the mixture of restriction fragments is fractionated on an electrophoresis gel, transferred to a nylon membrane, and placed in hybridization solution with a radioactive probe; the position of the fragment can then be determined by autoradiography. To avoid using radioactivity, probes can be constructed that can be detected by enzymatic methods. This provides a visible product that indicates the position of the probe on the membrane.

RFLP differences between plants are inherited in the same fashion as conventional genes, and genetic maps of RFLP's can be constructed using conventional methods. Such RFLP maps indicate the location of specific restriction fragments of chromosomal DNA relative to one another.

Other methods of detecting differences in genomes at the DNA level include those using the polymerase chain reaction (PCR). PCR is a technique for enzymatically producing many copies of a given portion of a DNA molecule

DNA cloning—*usually referred to as molecular cloning, which involves placing a piece of desired DNA in the chromosome of a vector organism that allows the DNA to duplicate or replicate during the division cycles of the host organism's DNA strand.*

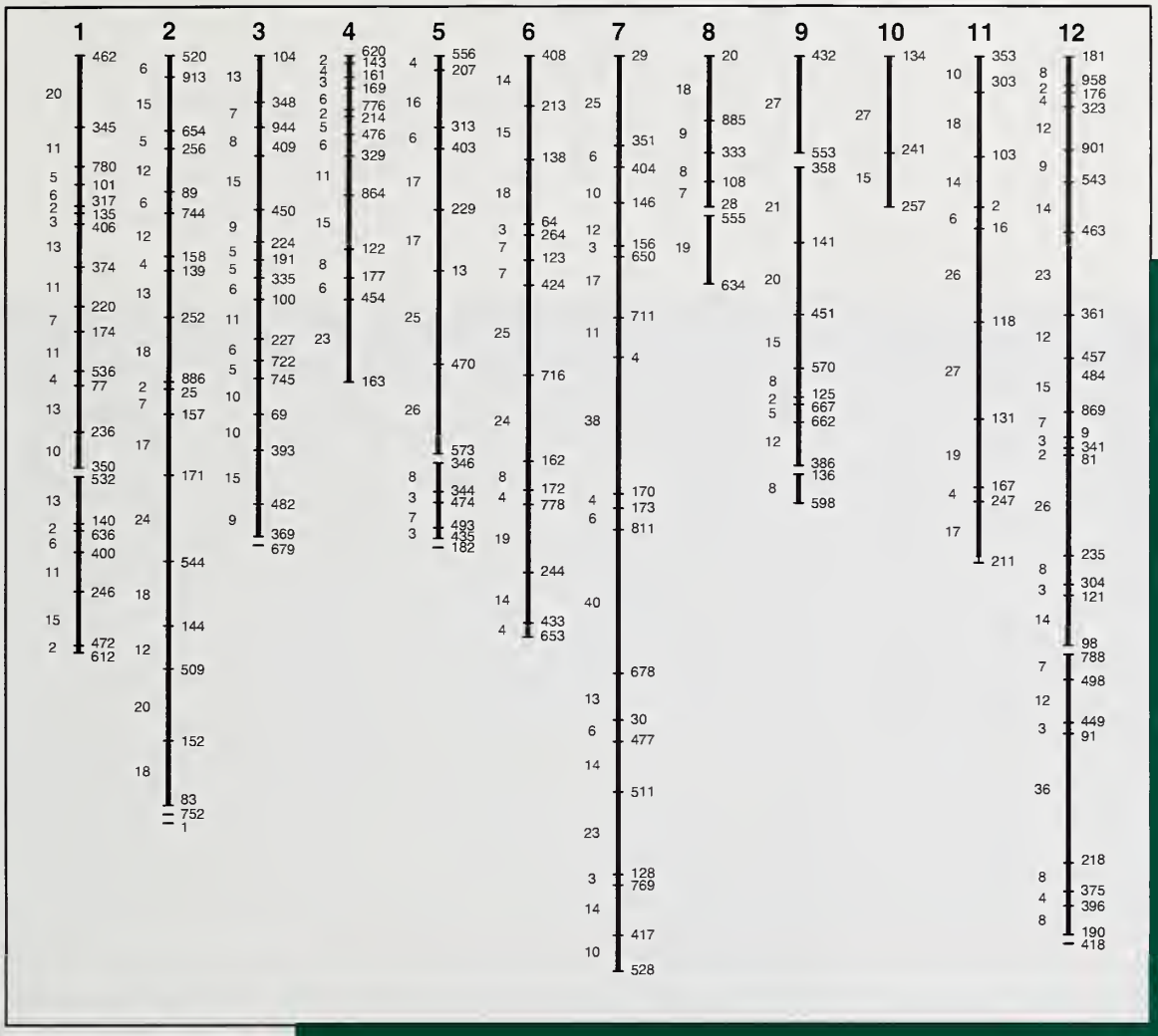
without the necessity of molecular-cloning the DNA. If the DNA sequence is known, oligonucleotide primers can be synthesized that complement the sequence. These need to be about 20-30 nucleotides long to ensure that they are unlikely to occur by chance at other locations in the genome. If the DNA template is denatured and the primer and nucleotide triphosphates added, DNA polymerase will be able to polymerize nucleotides onto the 3-prime end of the primer and faithfully copy the template. The newly synthesized product accumulates in a linear fashion. If, however, one simultaneously uses



Inheritance of RFLP markers compared with the inheritance of a gene for flower color.

a pair of primers complementary to opposite strands of the DNA template, and the distance between the primers is not too great (up to about 5 kilobase pairs), the synthesized product duplicates exponentially—the product synthesized with one of the primers has now become a template for the other primer.

Fragments of DNA produced by PCR amplification can be directly observed by staining after electrophoresis in an agarose or acrylamide gel. No cloning is necessary, and radioisotopes, DNA hybridization, and autoradiography are not necessary. So a large amount of data can be gathered in a short time.



An RFLP map of rice. The symbols represent the location of cloned DNA fragments used as RFLP probes.

PCR products can also be used as genetic markers if some difference between the fragments produced from different parental plants can be detected. Such differences could be of several types. If an insertion or deletion has occurred in the area amplified, the product may be of different lengths in the two parents and this difference may be detected directly. In some cases a product may be amplified from one parent, but the same primers fail to produce an amplification product from the other parent. Failure to amplify may be caused by a large insertion, which forces the primers to be too far apart for successful amplification, or by a DNA sequence change in the sequence complementary to one of the primers, which destroys the ability of the primer to initiate DNA synthesis. Changes in the area complementary to the 3-prime end of the primers are most likely to cause the amplification to fail.

In some cases, DNA sequence changes can be detected in the amplified region between the primer sequences if the amplification product is digested with restriction enzymes. Changes in DNA sequence occurring in restriction sites for a given enzyme can be detected by comparing the variation of restriction fragments produced after restriction enzyme digestion of the PCR products. Usually enzymes that detect four base pair sites are used, since these sites occur more often, and the method is sometimes called four-cutter analysis. Basically, it is RFLP analysis of PCR products.

The use of molecular markers has accelerated scientists' abilities to sift through the genetic complexity.

The main difficulty with using PCR products as genetic markers is that DNA sequence information is required to construct the primers. With regular PCR, a nonspecific product is unlikely to be generated—primers are intentionally constructed so they contain enough nucleotides that a complementary sequence is unlikely to occur by chance in a genome.

The use of a shorter primer, one made up of 8-10 nucleotides, presents the opportunity that several sequences complementary to it would occur by chance in a genome (a random sequence of 8 nucleotides would be expected to occur about 15,000 times in the tomato genome, assuming a random base composition). If, by chance, a pair of such sequences complementary to the primer occur on opposite strands of the DNA double helix within about 5 kbp of each other, the single primer would allow the production of a PCR product.

Large numbers of such random primers can be rapidly screened on the plants of interest, and those that produce clear amplification products (called Random Amplified Polymorphic DNA's or RAPD's) can be selected. PCR products resulting from such single random primers can produce genetic markers through the PRC products.

Again, as in all genetic mapping, it's necessary to detect some difference in the products produced in the parents of a cross. Short insertions or deletions in the DNA between the primers sites can be detected as length polymorphisms of the products, but most differences are detected as the presence or absence of a given amplification product. With the use of short primers, amplification is very sensitive to single base changes in the sequence complementary to the primer. If such a change occurs, the primer will typically not be able to initiate DNA synthesis, and no product will be formed. So RAPD markers are usually scored as dominant-recessive or plus-minus markers. The inability to distinguish heterozygotes from homozygotes means that less information can be derived from certain types of mapping crosses, such as F_2 populations. But RAPD primers can be easily produced and rapidly screened; they require no prior DNA sequence information, and a set of primers will work in many different organisms.

Molecular markers have several characteristics that make the construction of RFLP- and PCR-based maps much easier than the construction of conventional maps.

Primer—*a DNA sequence, generally short, that is paired with a DNA strand and provides a 3'OH endpoint or terminus at which DNA polymerase can initiate synthesis of a deoxyribonucleotide chain.*

Advantages of Molecular Markers as Genetic Markers

1. Natural occurrence. DNA nucleotide sequence variation is common in most organisms, because many nucleotide changes do not lead to important changes in genes. Therefore, many differences are found to be segregating in most genetic crosses. No mutagenesis needs to be performed.

2. Mapping in single crosses. Since a virtually unlimited number of markers will be found to be segregating in any one cross, a reasonably complete genetic map can often be constructed from the progeny of a single cross.

3. No effect on the organism. Since the sequence changes responsible for molecular markers seldom cause any change in gene products, they have no effect on the plant's form or function. Also, one marker has no effect on another; they segregate independently in a cross. In contrast, the mutants of conventional genes, which are often used for conventional mapping, usually have drastic effects on the plant and can interact in complex ways that make genetic mapping very difficult.

4. Constant nature. Molecular markers are scored using DNA samples isolated from the plant. One of the fundamental concepts of modern biology is that DNA does not change qualitatively during development of an organism. Therefore, scoring markers from DNA extractions from any part of a plant or any developmental stage is possible. It is not necessary for the plant to mature to a certain stage, as it might be to score genes such as those for grain characters.

A cornucopia of American agricultural bounty.

These, then, are the nuts and bolts of today's genetic research.

But let's briefly consider the trait in the plant, the farmer in the field. Now that science is able to locate specific genes precisely on chromosomes, we are entitled to look at the utility of this information and ask ourselves some fascinating "what if" questions.

What if important traits were pinpointed on that confounding double helix?

What if it were indeed possible, jigsaw-puzzle-like, to determine the locations of agronomically important traits relative to one another?

The practical implications are most alluring. If it were possible to map out important parts of specific plants' encoded genetic legacy, plant breeders might be able to more efficiently combine valuable genes to produce new plant varieties that are ideally suited for specific purposes in agriculture—designer plants such as:

- *a wheat resistant to rust fungi*
- *a soybean that simultaneously gives us high protein and high oil content*
- *a cotton plant with bolls that contain fibers of nearly uniform strength and length*
- *a corn plant that can withstand high temperature and drought yet still maintain productivity*

Physical Maps

Already we are able to determine how far apart on a DNA molecule two molecular markers may be. We can do this by using one of several molecular biology techniques.

To construct a physical map, it is necessary to measure the physical distance, which is proportional to the number of nucleotide pairs, between two DNA markers on the genome of the organism in question or to actually determine the sequence itself. Distances between markers on a physical map are always directly related to the number of nucleotide pairs in the DNA between the markers. Indeed, distances on physical maps are usually expressed in terms of thousands of nucleotide pairs (kilobase pairs, kbp) or millions of nucleotide pairs (megabases, mbp).

Several methods are used to measure physical distances between markers. One of these is electrophoresis in agarose or polyacrylamide gels, which can separate DNA molecules according to their size. Comparing the migration of appropriate molecular weight standards allows estimation of the number of nucleotide pairs present in a DNA molecule. Another method uses an electron microscope to directly visualize DNA molecules and measure their length. The length of a DNA molecule is proportional to the number of nucleotide pairs, and therefore physical mapping data determines results from comparison to standards included with the fragments to be measured. The ultimate measure of distance in nucleotide pairs is the actual nucleotide sequence of a DNA molecule.

Restriction Maps

Physical maps may be produced by various methods. For example, restriction maps are one common type of physical map. To construct a restriction map, digestion of the DNA of interest with one or more restriction enzymes produces a set of definite-sized restriction fragments. Since restriction enzymes cleave DNA at specific base sequences, the sequence at each end of these fragments is known. The enzyme EcoRI, for example, would produce a set of fragments, each of which has the sequence GAATTC (the EcoRI restriction site) at each end. To complete the restriction map, it's necessary to measure the size of the restriction fragments by gel electrophoresis and to determine the order with which they occurred in the parent molecule.

Restriction maps of an entire genome have only been constructed for organisms with small genomes, because large genomes produce so many restriction fragments that measuring and ordering them is very complex. Millions of restriction fragments would be produced by digestion of maize DNA with EcoRI, for example. Two general methods are being used to construct restriction maps of larger genomes. One involves the use of restriction enzymes that cleave DNA less frequently and produce fewer (but larger) restriction fragments; the other seeks to subdivide the genome into smaller units, to map each smaller unit, and then to order the units to arrive at a complete map.

These tomato plants have been genetically engineered to include phytochrome from oats. The plants are dwarfed, but bear normal-size, highly pigmented fruit.

But constructing these restriction maps is a mammoth undertaking, even when studying organisms that have comparatively small genomes. And while there are methods that simplify the job, there's a cost to this in terms of precision.

Restriction enzymes that cleave less often can be obtained by selecting those that have more nucleotides in their recognition site or those whose recognition site contains nucleotide sequences that are uncommon in the target DNA. Restriction enzymes that have eight-base-pair recognition sites, such as NotI, usually produce larger fragments than those that recognize six base pairs, such as EcoRI. A random 8-base-pair sequence would be expected to occur about 1 time every 65,000 base pairs as compared to 1 time every 4,000 base pairs for a 6-base-pair sequence, assuming random distribution of nucleotide pairs. So fewer, and larger, restriction fragments would be produced by NotI than by EcoRI. The genome of the bacterium *E. coli*,which contains a total of 4 mbp, has 23 NotI sites, and a restriction map for this enzyme is complete. The enzyme AvrII has only a six-base-pair recognition sequence, but part of this sequence occurs at a frequency much lower than would be expected on a random basis. Only 13 AvrII sites are present in the *E. coli* genome, and a complete restriction map of these sites is available.

Electrophoresis of large DNA fragments, such as those from rare-cutting restriction enzymes, requires modification of the conventional electrophoresis technique used to fractionate smaller DNA molecules. DNA molecules up to about 20 kbp will fractionate according to their molecular weight by conventional electrophoresis, but larger molecules move very slowly, if at all, through conventional gels and would not be separated according to molecular weight. Conventional electrophoresis uses a static electric field; DNA, which is acidic, moves toward the positive electrode. By periodically changing the electric field, one can force large DNA molecules to reorient so that they fractionate successfully.

Pulsed field electrophoresis is a general term for the several variants of the basic method. The NotI and AvrII restriction maps of *E. coli* were produced using pulsed field electrophoresis to fractionate the large fragments produced by digestion with NotI or AvrII. Since even the smallest crop plant genome is a hundred times larger than the *E. coli* genome, thousands of NotI sites would be expected to occur. Constructing restriction maps involving thousands of fragments is beyond the present capabilities of pulsed field gel systems, but such maps could be constructed for individual, smaller chromosomes, or large cloned fragments.

Contig Maps

Cloning small fragments of DNA has also rendered information that advances the process of gene mapping.

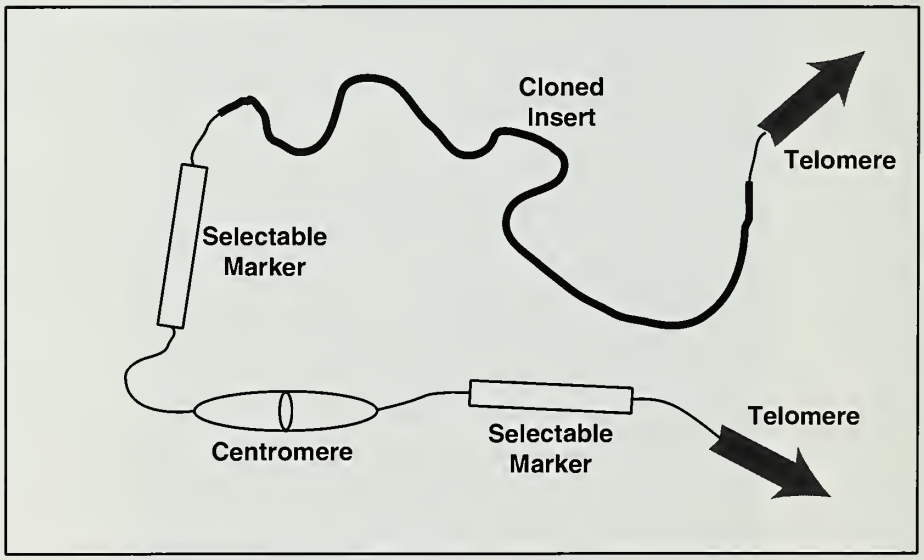
Another approach to producing complete physical maps attempts to clone the entire genome and then to determine the order in which the clones were present in the target genome. For this method to be completely successful, it is first necessary to construct a genetic library containing overlapping clones representing the complete genome of the organism to be mapped. The regions where two clones overlap will have identical nucleotide sequence and can be used to link the clones together to produce "contigs."

Overlap can be detected by making restriction maps of each of the clones and matching them to determine the order. Contig mapping has been attempted for several organisms, most notably the nematode *Coenorhabditis* and the plant *Arabidopsis*, both of which have small genomes. The limits of technology circumscribe the success of this approach, however. The main problem is that cloning vectors clone only relatively small DNA fragments.

The bacteriophage lambda contig mapping vector can clone DNA fragments up to about 40 kbp. This means that the number of clone arrangements that would have to be arranged into contigs is very large, 100,000 or more for a large genome. Another problem is obtaining a library that really contains all of a plant's genome. In the studies to date, certain regions are unclonable, probably because DNA clones are not stable in *E. coli* vector replication.

Yeast Artificial Chromosomes

One recent advance that promises to greatly aid the construction of contig maps is the development of new vectors for cloning of large DNA fragments One type of such vectors are yeast artificial chromosomes (YAC's).



A yeast artificial chromosome (YAC) vector.

YAC's capitalize on the fact that chromosomes of all eukaryotic organisms appear to have certain essential DNA sequence elements in common. These include centromeres, which regulate movement of chromosomes during cell division; telomeres, which stabilize the ends of the chromosomes during DNA replication; and autonomously replicating sequences (AR's), which serve as initiation points for DNA replication. Virtually any piece of DNA will behave as a chromosome if it possesses these three elements and can be placed in a yeast cell.

DNA clones from 200 to 500 kbp or more are possible in YAC's. YAC's are finding two main uses in contig mapping. One is to bridge the gap and link existing contigs of lambda clones; the other is through the construction of contig maps entirely with YAC clones. Since larger fragments can be cloned with YAC's, fewer overall clones are necessary to complete a map.

Physical Mapping Markers

A map requires markers to specify position and relative location. Recombination genetic maps use genes that code for morphological and physiological traits of the plant, or RFLP marker clones that follow Mendelian genetics. Physical maps use different sorts of markers that depend on production procedures.

Restriction maps use restriction sites as markers, but one could also use the location of cloned DNA sequences, such as RFLP markers. DNA clones have some disadvantages as mapping markers, however. Someone must maintain the clones and distribute them to prospective users. With the thousands of clones that will be necessary for any mapping project, some are sure to be lost or mislabeled.

Sequence Tagged Sites (STS's) are an alternative to cloned sequences as mapping markers. STS's have become possible because of the use of the polymerase chain reaction. PCR allows the selective amplification of a region of DNA up to 10 kbp, providing that the sequence of a 20-30 nucleotide area flanking the area of interest is known. Primers complementary to the flanking sequences are made in a DNA synthesizing machine for use in PCR. A sequence tagged site is a set of two complementary primer sequences that will faithfully amplify a specific region. Published sequences of sequence tagged site markers offer researchers the opportunity to produce this marker by PCR technology. This reduces or eliminates the necessity to maintain clones. Therefore, the DNA information is less likely to be lost or mixed up.

DNA Sequence Maps

The ultimate physical map would be the DNA sequence of an entire genome, and such a map supersedes all other types of physical maps. Technology is presently the factor limiting the construction of complete sequence maps, making it too time-consuming and expensive to completely sequence any but the smallest genomes, such as viruses. Improvements in automation of sequencing and in the scale with which sequencing can be done will have to be accomplished before DNA sequence maps of large plant genomes will be possible.

So, while scientists are energized by the futuristic promise of complete DNA sequences for an entire genome, they're tuned in to the need to set attainable goals that best maximize finite research dollars and resources.

Inspecting a film autoradiograph.

Genome Mapping Applications

Present-day scientific realities suggest a strategic course of priorities for federally funded plant genome research efforts. The following pages outline some urgent production issues in agriculture that can be addressed by genetic research.

Crop production limits are becoming evident in some major crops. Some limitations are intrinsic to the plant's own genome and its interaction with the physical environment. Others represent challenges to the well-being of the plant by other organisms. Intrinsic limitations include such factors as total yield, amount and amino acid composition of seed proteins, resistance to wind and rain damage, and ability to grow under extreme conditions of temperature, of salinity, or in marginal soils. Biological agents that limit crop production are primarily insects, fungal, bacterial, or viral pathogens, and competing plants, which we call weeds.

Modern monoculture agricultural methods increase the biological pest problem. To facilitate mechanical planting, harvesting, and processing on a large scale, breeders have produced crops that have uniform-sized seeds that synchronously germinate and produce a field of plants with uniform stature and structural properties. Simultaneous maturation is necessary for mechanical harvesting, and the product needs to be uniform because of mechanical processing or consumer preferences. Since all these plant properties are genetically controlled, there is continuous pressure to narrow the genetic base, to produce crops with less genetic variability. Lack of genetic diversity promotes catastrophic epidemics of insects or pathogens and, in fact, provides a strong selection pressure that tends to promote the appearance of new pathogen biotypes.

The struggle between agriculturalists on one side and insects, pathogens, and weeds on the other has often been compared to an arms race in human society. Plant breeders strive to produce new varieties that can be grown in large scale but are resistant to all insects and pathogens. Varieties that are not resistant are protected by treatment with insecticides or fungicides. Elimination of competition from weeds is increasingly being controlled by herbicide applications, and often multiple applications are necessary each growing season. New plant varieties do well when first released, but evolution has equipped insects and pathogens with many ingenious and powerful weapons to overcome plant resistance, and weeds often evolve resistance to herbicides. Then another round in the arms race is necessary to provide new resistant varieties, and the cycle continues.

To produce new plant varieties, it's necessary to change the genetic makeup of the crop in question. Desirable genes have to be incorporated into the crop, and undesirable genes have to be eliminated or replaced. In other words, one needs to genetically engineer the plant to meet the demands of agriculture. Genetic engineering of crop plants necessitates methods of identifying potentially valuable genes and then transferring these to the crop that one desires to improve. Agriculturalists have been practicing genetic engineering for thousands of years by conventional crossing and selection. The principal limitation to genetic engineering of plants by conventional plant breeding has been technological; transferring desired genes from one plant to another has been a very laborious process with serious limitations.

Until recently, the only method of introducing genes into a plant was by crossing it with another plant containing the desired gene or genes. A plant breeder searching for desirable genes would have to find such genes in plants that could successfully cross with the crop in question. The genes available to the plant breeder for any crop (the gene pool) would be limited to closely related plants, because these would be the only ones that could cross with the crop plant. Transferring genes by sexual crosses has other limitations. The product of a genetic cross receives half its chromosomes and hence half its genes from one parent and half from the other. If a crop plant is being crossed with a wild relative, for example, this almost always results in the incorporation of large numbers of undesirable genes along with the few desired genes. Further rounds of crossing and selection are necessary to eliminate undesirable genes while retaining the "good" genes. The only way to select plants containing desirable genes is to plant a large number of plants from a cross and to observe the effect of the genes on the plant's morphology, its physiological characteristics, its resistance to an insect or pathogen, or some other aspect of the plant's phenotype.

Selection is often a rate-limiting step in plant improvement. This is mainly because no way has been available to directly select for the presence of the DNA that makes up a gene or for the primary products of gene action. Instead, plant breeders have had to depend on observing the final products of gene action. Since gene action is often specific for certain developmental stages or tissues, plants may have to be grown to maturity to select genes influencing the final product, such as seed yield. Special enclosures may have to be constructed and insects cultured to test insect resistance. Other genes may influence or mask the expression of the gene that the plant breeder is trying to select, particularly in early generations of a cross between distantly related plants.

Recently developed molecular genetic techniques present the plant breeder with a new set of tools to attack traditional plant breeding problems. Still, it's important to realize that the challenges facing plant breeders are the same. Plant breeders will still have to screen the gene pool for valuable genes for introduction into crop plants, to actually move these genes into the crop, and to evaluate the performance of the newly engineered crop.

But the new tools of biotechnology will provide a more powerful way to accomplish the basic goal. First of all, the gene pool available to plant breeders will be expanded to include virtually all organisms, plants, animals, bacteria or viruses, because of techniques of gene cloning and transformation. Cloned genes introduced into plants by transformation can be directly selected at the DNA level, and their expression can be monitored by direct detection of primary gene products. Even genes that have been introduced by conventional crossing methods can be selected by linkage to RFLP markers, making selection independent of gene expression. Molecular genetic mapping and DNA sequencing techniques will make it possible to locate genes and to clone them without first having to characterize their gene products.

All these new tools require increased knowledge about plant genomes and new techniques for obtaining, storing, and using this information.

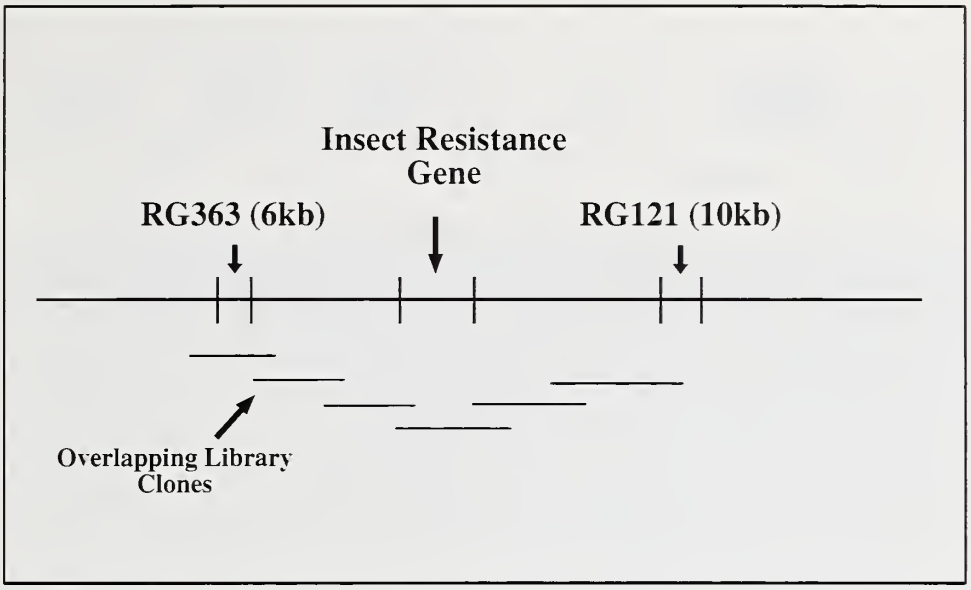
RFLP-Assisted Plant Breeding

Genetic maps based on molecular markers, such as RFLP's, are general purpose tools with a variety of plant breeding applications. The first step in the use of RFLP maps for conventional breeding is to construct a map with RFLP markers distributed at fairly regular intervals over each of the chromosomes. Each gene will then of necessity be fairly close to (linked to) one or more RFLP markers. By analyzing segregation patterns of both the gene and the RFLP markers in crosses, cosegregation can be used to determine which RFLP markers are linked to the gene. A gene found to be closely linked to an RFLP marker is said to be "tagged" by the marker.

Tagging allows plant breeders to use indirect selection for genes of interest. Instead of trying to select which progeny plants contain a gene of interest by checking for the action of the gene, researchers extract DNA from the plant and check for the presence of the linked RFLP marker.

Very small plants or seedlings can be used for DNA extraction, and a large number can be rapidly screened for the presence of the desired gene and a minimum amount of chromosome material containing undesirable genes. Indirect selection could fail if recombination from crossing over were to occur between the linked RFLP marker and the desired gene, but even low-resolution RFLP maps give a high degree of accuracy. For example, if a gene is located between two RFLP markers and is 10 map units from each marker, indirect selection using both RFLP markers would work about 99 percent of the time.

Savings in time and expense from RFLP-assisted plant breeding can be considerable. In a conventional back-crossing program, for example, many generations of crossing can be avoided. Once a gene is tagged with RFLP markers, it can be easily moved to other varieties by indirect selection of the progeny from the appropriate cross.



Chromosome walking from linked RFLP markers.

Expanding the Gene Pool

A few crop plants can be directly transformed, and genes can thus be added to them without going through the crossing process. Much current research is being carried out to make these processes more efficient and to increase the number of crops that can be transformed. Ability to transform plants opens up a whole new world of possibilities for plant breeders.

Potentially valuable genes from virtually any organism can be transferred into transformable plants, so the gene pool available to a breeder becomes unlimited. A backcrossing program, for example, that might take 8 to 10 generations could be accomplished in a single step. But basic knowledge of crop systems is still absolutely required to identify factors limiting plant production, to suggest potentially valuable genes, and to evaluate the resultant transgenic plants.

Discovering agronomically valuable genes in a gene pool can proceed in two main ways. One of these is the direct approach. If it is desired to find genes for resistance to a certain fungal pathogen, for example, one way to find them is to plant a variety of plants, inoculate them with the fungus, and observe the result. If the fungi fail to infect the plant, genes for resistance to the fungus may be present. The other way to find genes that might confer fungal resistance would be to study the details of the plant-pathogen interaction to discover ways in which infection could be thwarted. Once the infection and plant defence processes are understood, it might be possible to design ways to enhance the plant's defenses or to inhibit some fungal processes crucial for infection. So far, most valuable genes have been located by the direct approach, and very little is known about the mode of action of these genes.

Map-Based Gene Cloning

In order to expand the gene pool available to plant breeders for use in transformation of crop plants, it is necessary to clone potentially valuable genes. Most genes have been cloned in a two-step process. First, a random genomic library is produced in the hope that every segment of the genome will be present in one or more clones. This is the easy part.

Since genomes of higher order organisms are so large, and cloning vectors can take only relatively small DNA fragments, the problem then becomes selecting the clone of interest from the hundreds of thousands of other clones in the library. This has most often been accomplished by constructing a DNA or RNA probe that would be able to recognize and distinguish the desired clone from all the others. If the RNA or protein product of the gene can be obtained, a probe can be constructed because the nucleotide sequence of the gene will be reflected in its product. Therefore, most genes that have been cloned have been those that synthesize a lot of gene product or produce the product only in certain tissues or developmental stages, making it possible to isolate the product.

But most genes of agricultural interest have been located by the direct method, and nothing is known about the gene product. In most cases the gene product will be present at low concentration, and discovering what it might be would require a great deal of time and effort. Genes can be cloned, however, without knowledge of the product. One way this can be done is by map-based cloning based on tagging with RFLP markers.

If genes can be tagged with closely linked RFLP markers, then sections of the genome that are close to the desired gene can be selected from a genomic library by using the RFLP markers themselves as probes. If the RFLP markers are very close and a library of very large fragments, such as YAC's, is available, the gene of interest might be on the same clone as the RFLP marker. This is unlikely, however, because it is difficult to obtain RFLP markers that are close enough. One will have to "walk" down the chromosome by selecting a clone adjacent to the RFLP marker clone, then the next adjacent clone, and so on, until the gene is reached.

Chromosome walking will be rendered superfluous when complete physical maps of genomes are available. If a complete nucleotide sequence map were available, the RFLP markers could be found in the sequence and all closely linked genes could be identified from the nucleotide sequence.

Quantitative Traits

If all the traits that confer agricultural value on a plant were each controlled by single genes, things would be greatly simplified for plant breeders. But most traits of agricultural interest are controlled by several genes. Yield, which might arguably be called the most valuable trait of all, is one of these polygenic traits. It's easy to understand why such a trait would be polygenic; many characteristics, such as water-use efficiency or photosynthetic efficiency, could contribute to the final yield. A complex trait such as yield is easily measured on a quantitative basis, but determination of the underlying genetic basis is very difficult. For most quantitative traits the number of genes, or Quantitative Trait Loci (QTL's), involved in the trait and the chromosomal location of those genes is not known. Since all the genes involved contribute to the same final measurable trait, it's difficult to determine the effect of each individual gene.

RFLP maps provide a way to determine the effect of the individual genes that combine to produce a quantitative trait. This is possible because the segregation of a large number of RFLP markers can be followed in a single genetic cross. If these markers have been mapped, it is possible to follow the segregation of every chromosome segment individually and to correlate the presence of a certain chromosome segment with the quantitative trait. In this way the number of chromosome segments, and hence a minimum estimate of the number of genes contributing to a trait, can be determined. Since the RFLP markers have been mapped, the chromosomal location of the quantitative trait genes is known. Using RFLP-assisted selection, plants can be assembled that contain several favorable genes for the trait and that do not contain unfavorable segments. This sort of analysis also paves the way for eventual map-based cloning of genes controlling quantitative traits.

Mature cotton boll at left was protected by a gene taken from *Bacillus thuringiensis*, whereas other bolls show damage from cotton pests.

Setting Our Research Priorities

We are on the threshold of a new era in crop production. Recent advances in genome research and biotechnology promise to greatly enhance our ability to achieve the basic goals of plant breeding and agriculture in general. To take advantage of this tremendous opportunity, additional work needs to be concentrated in several areas of genome research.

Additional Maps

Maps of conventional genes are available for only a handful of crop plants, and their use in plant breeding is limited. Low-resolution RFLP maps can be rapidly constructed and are useful in a wide variety of applications, so maps of all important crops should be constructed.

New Technology

How efficiently and accurately a job can be done is always a reflection of the tools available. To gain the ultimate benefits from mapping of plant genomes, new tools require development, modification, and refinement. In many cases the best system for development of tools is not the final system to which the tools will be applied. Use of model systems favorable for tool development will sometimes be required. Some of the main areas that will require new tools and approaches include the following.

Dealing with Larger DNA Fragments

As has been outlined above, the ability to construct contig maps and to do chromosome walking depends strongly on being able to manipulate large DNA fragments. New and improved cloning vectors and analytical procedures for large DNA fragments beg for exploration.

Improved DNA Sequencing Methods

Present technology that allows the sequencing of only short DNA fragments is very expensive in terms of labor and supplies. We need automatic sequencing methods at a fraction of the cost of present techniques.

Identification of Valuable Genes

More basic knowledge of the processes involved in plant growth and development, insect and disease resistance, and response to environmental variables needs to be gained before advances can be made in most areas.

Information Storage

A large amount of information is rapidly being accumulated on DNA sequences, molecular markers, and screening for various traits. As our tools improve, the rate of information accession will greatly increase. This information would overwhelm presently available systems for its storage, retrieval, and dissemination.

Exploring a Significant Challenge

Plants have many advantages for basic genome research. Large numbers can be produced and analyzed, controlled crosses can be made for genetic analysis, single cells can often be transformed and regenerated to entire plants, and there are few moral or ethical problems in experimentation with plants. Plants can thus contribute to basic advances in all fields of biology. The next few years will be exciting times. Plant genome research will be an intellectually exciting enterprise that should attract our best scientists, and the plant genome promises to reveal many wonderful surprises.

Plants sustain all life on Earth, and it is difficult to overestimate the importance of gaining additional basic and applied knowledge about them. The very future of the human race depends on it.

NATIONAL AGRICULTURAL LIBRARY
1022946836